Journal of Research on Educational Effectiveness, 8: 245–279, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2013.878011

METHODOLOGICAL STUDIES

Adding Design Elements to Improve Time Series Designs: No Child Left Behind as an Example of Causal Pattern-Matching

Manyee Wong
American Institutes for Research, Chicago, Illinois, USA

Thomas D. Cook
Northwestern University, Evanston, Illinois, USA, and Mathematica Policy Research, Inc., Washington, DC, USA

Peter M. Steiner
University of Wisconsin–Madison, Madison, Wisconsin, USA

Abstract: Some form of a short interrupted time series (ITS) is often used to evaluate state and national programs. An ITS design with a single treatment group assumes that the pretest functional form can be validly estimated and extrapolated into the postintervention period where it provides a valid counterfactual. This assumption is problematic. Ambiguous preintervention functional forms are common, as are other factors affecting posttest means and slopes. Using No Child Left Behind as an example, we demonstrate how adding multiple design elements to the basic ITS structure serves to promote causal inference by limiting alternative interpretations. No added design element is perfect by itself, but we argue that they *collectively* provide a strong causal warrant when the predictions they engender are complex, the results "cohere" with the predictions, and no alternative can fit the same pattern of predictions even if it can fit some of them.

Keywords: No Child Left Behind, pattern matching, coherence, comparative interrupted time series

INTRODUCTION

The One-Group Interrupted Time Series Design

Sometimes it is important to evaluate objects that do not lend themselves to randomized control trials (RCTs), regression discontinuity designs (RDD), or even sophisticated matching techniques. One salient example of this is when national or state programs are under scrutiny, many of which entail high financial costs, gain a high political profile, apply to

Address correspondence to Thomas D. Cook, Institute for Policy Research, Northwestern University, Evanston, IL 60208, USA. E-mail: t-cook@northwestern.edu

the entire population, are expected to have broad impact, and cannot be denied to those eligible. Creating a credible comparison group can then be difficult and often leads to the use of interrupted time-series (ITS) designs.

The most basic ITS design involves a single treatment group that, at a known time, receives the intervention under study and is then repeatedly observed on the study outcome both before and after the treatment. The time series that results can involve the same persons over time, as in a true longitudinal design; or it can involve repeat cross-sections as with successive cohorts of respondents from a particular population; say, fourth graders within a school district. The statistical analysis of the ITS design requires (a) accounting for the likely nonindependence between observations and (b) estimating the preintervention temporal functional form and extrapolating it into the posttreatment period where it provides the counterfactual against which the obtained postintervention means and slopes are compared. The resulting analysis involves a single difference—between a short extrapolation and the obtained immediate posttest mean or else between a longer extrapolation and the obtained posttest slope.

But not all pretest trends are well estimated. Moreover, the simple ITS design can also run afoul of other threats to internal validity (Cook & Campbell, 1979). One is *statistical regression,* caused when the intervention is introduced as a response to observations that suddenly shift prior to the intervention. Another is *history,* due to outcome-correlated events that are not related to the treatment but that occur between the intervention onset and first indications of a possible effect. A third is *selection,* due to a postintervention change in the composition of the units providing outcome data. A fourth is *maturation,* due to an inflection in the rate of maturation (or secular change) that occurs around the intervention point. The final one is *instrumentation,* a postintervention change in how the outcome is assessed.

Some applications of the one-group ITS design turn out to be causally interpretable. Shadish, Cook, and Campbell (2002) presented some examples. But all are characterized by the availability of short time intervals, stable preintervention time trends, immediate and very large impacts, and no visual evidence of a preprogram dip. Much more typical in actual research practice are fewer time points, longer time intervals between data points, uncertain preintervention functional forms, and delayed effects that increase the odds of viable alternative interpretations based on history and selection. Something better is needed than a simple ITS design.

Adding a Single Comparison Time Series to the One-Group ITS Design

To reduce the number And viability of internal validity threats, a single no-treatment comparison time series is often added to the one-group ITS, thus creating a comparative ITS design that can be abbreviated as a CITS design. A special kind of difference-in-differences design (DD) analysis is then called. But it is more complex than most DDs that contrast just four means (see Angrist & Pischke, 2009; Lechner, 2010). Thanks to the multiple preintervention measures it is now possible to reliably assess mean differences between the treatment and comparison groups *over multiple times,* and these initial *slope differences* can serve as the causal counterfactual for estimating changes in mean or slope from before to after an intervention. Combining the treatment and comparison time series means that any potential history, selection, maturation, instrumentation, and regression effects in the comparison time series can now be differenced out from the mean and slope changes observed in the treatment series.

However, such differencing cannot account for threats operating *differentially* across groups, for example, local historical forces that differ between the treatment or comparison time series near intervention time, or regression artifacts that operate more in one group than the other at that time. Such differential threats are usually less plausible than the more general internal validity confounds threatening the one-group ITS design, but they are far from impossible. As useful as it is to add a comparison time series, something more is preferable.

Adding More Design Elements Than Just a Comparison Group

Generic Purpose

This article illustrates the advantages of adding other design elements to the basic ITS in addition to a comparison series. Design elements are structural research features, many identified by Fisher (1925, 1935), that improve causal counterfactuals without necessarily making them “perfect.” Such design elements include many features of (a) *Comparison Groups,* such as how many groups are used in a given study, how initially equivalent they are to a treatment group, and how treatment and comparison cases are matched on key attributes correlated with both selection and outcome; (b) *Measurement Attributes,* of which probably the best known are the number of pretest time points and use of “non-equivalent dependent variables” that should not be affected by the treatment but should be affected by other alternative interpretations; (c) *Treatment Allocation Attributes,* where the best known are treatment assignment by a random process, or by an otherwise fully known selection process (as in RDD), or by a partially known selection process that is supplemented in some way; and (d) *Repeated Treatment Applications*. This can involve introducing, removing, and even reintroducing a treatment within a single study, doing this either with a single treatment group or at a later date in the original comparison group.

We soon describe a CITS application that allows many tests of treatment/comparison differences in differences in mean, slope, and the two combined. These multiple tests are made possible by the deliberate use of two versions of the treatment, three comparison groups, two data sets, and three study outcomes. Together, these design elements enable us to describe a unique pattern of data that aids causal interpretation because (a) no alternatives predict the same pattern, even if they do predict parts of it, and (b) the obtained data correspond with the predicted complex pattern. However, meeting these unique prediction and empirical correspondence conditions is not easy, and no claim is made here that testing a complex pattern of predictions is as good as random assignment or regression discontinuity.

Theory of Inference

The epistemological warrant for asserting that empirically corroborated complex predictions facilitate causal inference in observational studies has a long history in statistics, psychology, and economics. However, the very idea seems at odds with recent causal thinking that emphasizes simple rather than complex causal hypotheses. Here are three salient examples.

When a valid random assignment procedure has been correctly implemented, an RCT can validly compare the difference between as few as two posttest means. RDD can also be conceptualized as a test of the difference between two means at the prespecified cutoff value where the two populations hardly differ. Propensity scores seek to mimic the logic of random assignment and share the same preference for testing simple hypotheses about the difference between two means, albeit matched means from which hidden bias may not be completely ruled out. In all three cases, the causal hypothesis can be simple in form because it is supported by a logical warrant that slightly varies by design type—that the two groups are identical in expectation (RCT), that they are almost identical and differ by only 1 point on the assignment variable (RDD), and they are matched on most of the covariates accounting for selection (propensity scores).

Contrast the simplicity of these such externally warranted causal tests with what Rosenbaum (2005) wrote about R. A. Fisher:

> Cochran (1965) summarizes the view of Sir Ronald Fisher, the inventor of the randomized experiment: About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate." The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold. (p. 1,459)

Fisher's point is that, if a given theory predicts a complex pattern of outcomes that is subsequently empirically corroborated, then it is all the more likely that no other theory can match this pattern; in the extreme, none might. Fisher noted another advantage of elaborate theories. Systematically obtaining a contingent pattern of results can help specify the boundary conditions over which a causal relationship varies, thus promoting external as well as internal validity. Fisher (1935) wrote, "Any given conclusion ... has a wider inductive base when inferred from an experiment in which the quantities of other ingredients have been varied. ... Standardization weakens rather than strengthens our grounds for inferring a like result, when, as is invariably the case, these conditions are somewhat varied" (p. 99).

Rosenbaum (2009, 2011) has extended Fisher's insights in his own work on "coherence." Coherence entails starting with a single causal force and using the theory associated with it to hypothesize a pattern of results that will collectively rule out, or render consensually implausible, contending theories that do not predict the same pattern even if they do predict some parts of that pattern. The next step is to test if the obtained data corroborate the unique multivariate pattern of predicted hypotheses. Rosenbaum stressed that this testing process is likely to involve some hypotheses that are statistically independent and others that are not. Moreover, the testing may focus on unique hypotheses from within the total set or on an omnibus test of the ensemble of hypotheses. These testing issues aside, the main theoretical concerns are for (a) the intellectual coherence of the multiple predictions/hypotheses emanating from the cause under scrutiny and (b) the empirical coherence of the fit between the data and these various hypotheses, however they are tested. Rosenbaum is also careful to note that that we cannot logically know all the competing theories capable of generating the same set of hypotheses. So empirical coherence cannot necessarily rule out all sources of hidden bias across all the hypotheses tested. His more modest claim is that a complex pattern of predicted and obtained results renders most alternative interpretations implausible.

Rosenbaum (2005) explicitly noted the similarity between some of his own work in Statistics and Donald Campbell's (1966) work on "pattern matching" in psychology. He wrote of some of Campbell's collaborators that "Cook & Shadish (1994, p. 565) [say]: 'Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations'" (p. 1,455). Campbell himself asserted that, in nonexperimental work, obtaining data that match a complex prediction rules out more alternative interpretations than do tests that match data to simple causal hypotheses like the difference between two or four means as in DD. Reflecting this, subsequent writing on pattern matching (Corrin & Cook, 1998; Shadish et al., 2002) has come to emphasize testing hypotheses specified as higher order statistical interactions rather than main effects (a mean difference) or first-order interactions (DD). Campbell (1966) acknowledged that, in actual research practice, it is not yet possible to match multivariate data to certain kinds of elaborate theory in any causally convincing way. He especially pointed to theories that postulate multiple mediated pathways within a temporal chain of causal influence, the type of endogenous theory that structural equation modeling is often used to test even though it fails to identify causal links that come later in such models. Not all ways of mapping multivariate data onto complex causal predictions are equal, in his view. Best is where a single causal agent is under analysis and it predicts multiple independent effects; the worst is when a set of temporally mediated and clearly endogenous relationships are being tested.

Details about elaborate theories, coherence, and pattern matching overlap with the call in economics to move beyond DD tests of two pretest and two posttest means. In his discussion of ways to improve quasi-experiments, Meyer (1995) concluded, "Design complications such as multiple treatment and comparison groups and multiple pre-intervention . . . observations are advocated" (p. 151). In essence, his call is to supplement the four means of the standard DD by adding pretest time series data, more than one comparison group, or both of these features in order to increase the stringency of tests of a single causal hypothesis. Even more radical are generalized difference-in-difference studies (GDD) in economics. Imbens and Wooldridge (2007) wrote of the simplest GDD:

> A more robust analysis than either of the DD analyses described above can be obtained by using both a different state and a control group within the treatment state. . . . The coefficient of interest is now the coefficient on the triple interaction term. . . . For obvious reasons, the estimator . . . is called the difference-in-difference-in-differences (DDD).

There are now many applications of the DDD strategy in Economics, for example, Chetty, Looney, and Kroft (2009) on tax policy; Solé-Ollé and Sorribas-Navarro (2008) on intergovernmental transfers; Milligan and Stabile (2009) on child benefit expansions; and Michalopoulos, Bloom, and Hill (2004) on how the bias from nonequivalent comparison groups in job training evaluations varies by whether the comparisons are or are not local to the treatment group. GDD studies with more than three differences are also available, including those examining the local economic effects of hurricanes that occur in up to a dozen different places at different times (e.g., Belasen & Dai, 2011; Belasen & Polachek, 2008; Ewing & Kruse, 2005). Whether a time series DD with multiple predictions is at issue, or a GDD, the requirements are the same: The obtained data have to match the predicted level of causal contingency, and no competing cause should account for all the contingencies even if it could account for some of them.

In none of these disciplinary contexts is mapping complex data patterns onto complex predictions sufficient for causal validation. Across the discussions, four limitations stand out:

1. Theoretically, it is not possible to identify every single competing theory, and any theory that escapes research attention might predict the same pattern of results as the theory under test.

2. Methodologically, sharp tests of multiple predictions require very stable data lest random error masquerades as failure to corroborate the expected coherence pattern.
3. Statistically, tests of the correspondence between complex predictions and obtained data have to account for any dependencies among tested hypotheses.
4. Although ITS helps with coherence tests, it is not necessary for them. Indeed, Corrin and Cook (1998) outlined some studies without time series data but with multiple design features nonetheless.

However, a treatment and a carefully constructed comparison series offer two of the strongest features in Fisher's repertoire of design elements, and that is probably why most of the pattern-matching examples in Shadish et al. (2002) include both of these features. So does the No Child Left Behind (NCLB) application we present next and about which we ask, Has it raised academic achievement nationally?

Our case is that causal inference is improved by adding to the basic ITS, not just a comparison series but other design features too. With regard to NCLB, we make 39 specific hypotheses designed around two different ways of conceptualizing the treatment; three different kinds of comparison time series; two data sets; three combinations of academic grade and subject matter; and DD tests of means, slopes, and means plus slopes. The purpose of so many comparisons is not replication per se; it is to ensure that no alternative interpretations remain unaddressed and to identify some boundary conditions over which NCLB's effects vary.

The Application: NCLB

Officially enacted in January 2002 and applying to all school districts that accept federal funds, NCLB requires each state to (a) conduct regular achievement testing using its own standardized tests that have been aligned to its own curriculum standards, (b) establish a clear time schedule by which an increasing fraction of students should become proficient by the state's performance standards, and (c) impose sanctions on schools that fail to make adequate yearly progress (AYP). The severity of these sanctions increases with the number of years a school fails to make AYP. One year of failure requires informing parents, whereas 5 consecutive years entails closing or restructuring the school (Goertz, 2005; U.S. Department of Education, 2008).

One feature of NCLB deserves special attention. The original program goal was for all public school students to be proficient by 2014. To this end, states had considerable freedom to create their own tests, set their own passing standards, and establish their own timetable for getting all students proficient by 2014. In NCLB's earlier years, states with easy tests, low standards, and back-loading the date for universal proficiency will have relatively few failing schools and so will have to undertake relatively little educational change. In contrast, states with more difficult tests, higher standards, and a front-loaded schedule should have more failing schools initially and so have to undertake more reform initially.

Many studies have tried to assess NCLB's merits. Most are narrowly focused on a particular aspect like the law's highly qualified teacher provision or its unintended consequences for teacher morale (Finnigan & Gross, 2007; Neal & Schanzenbach, 2010; Rosenberg, Sindelar, & Hardman, 2004; Smith et al., 2005). Few studies have evaluated the overall impact of NCLB on student achievement. The congressionally mandated National Assessment of Title I (Stullich, Eisner, & McCrary, 2007) presented time series data from National Assessment of Educational Progress (NAEP) to describe achievement trends before and after NCLB; but it counseled against causal interpretation, offered no statistical analyses, and it did not describe the achievement-correlated forces co-occurring with NCLB. The Center on Education Policy (2007) used state trend data between 2002 and 2005, claiming that the percentage of proficient fourth- and eighth-grade students increased by 1 to 3 percentage points in a majority of states. However, no comparison group data were presented, and, thus, many internal validity threats were not ruled out. Fuller, Wright, Gesicki, and Kang (2007) suggested that growth on the fourth-grade NAEP reading test faded after NCLB and slowed for math after 2002. Lee (2006) compared states with high-stakes testing and strong accountability systems prior to NCLB with states that only adopted a stronger accountability system after NCLB, finding no effects. Dee and Jacob (2011) conceptually replicated Lee (2006), contrasting states whose accountability systems had no links to sanctions until NCLB (the treatment time series) with states whose accountability standards were already tied to sanctions pre-2002 (the comparison time series). They detected a significant fourth-grade math effect. But due to NCLB, the comparison states with sanctions before 2002 might still have modified their sanctions after 2002, the study design precludes a national estimate of NCLB's impact, and only part of NCLB was evaluated anyway. The authors write, "Our estimates will capture the impact of the accountability provisions of NCLB, but will not reflect the impact of other NCLB provisions such as Reading First or the 'highly qualified teacher' provision" (p. 427). So Dee and Jacob may have underestimated the total impact of NCLB on public schools. To calculate this requires more than just a contrast between public schools in some states versus others; also needed is a contrast between the nation's public and private schools.

The Design Elements

This article presents a variety of different design elements added to the basic ITS design in order to rule out alternative interpretations and to identify causal conditionals.

Two Treatment Variants and Three Nonequivalent Comparison Time Series. We specify three contrasts, noting and justifying the treatment variant and comparison intrinsic to each.

Contrast 1: Public versus Catholic schools. NCLB applies to any school district accepting federal funds, and thus almost exclusively to public schools. Catholic schools do have a minimal involvement, though. After 2002, the two most popular NCLB programs in Catholic schools were (a) Reading First, with roughly 3% of Catholic school students participating, and (b) Title 1 Part A, in which 6% of students participated (U.S. Department of Education, 2007). In total, only 4.7% of all K-12 Catholic school students receive any type of Title I services (Keigher, 2009). But even in the rare Catholic schools to which NCLB applied, it did not do so in the same way as in public schools. Catholic schools do not have to use test scores to determine whether to make changes in school practices, though some may do so. Nonetheless, given NCLB's low coverage in Catholic schools we treat Contrast 1 as very close to a contrast of NCLB present versus absent.

The suitability of Catholic schools as a stand-alone no-treatment control group is somewhat compromised because of publicity about priest sexual abuse that emerged in 2002. This is one reason for the other contrasts we develop. Even so, there is scant evidence that the Catholic school losses could have had much of an effect on public schools. Table 1 provides historical data on enrollment in public, Catholic and non-Catholic private schools. Public schools account for about 90% of national enrollment, and Catholic and non-Catholic private schools are each responsible for about 5%. Catholic schools have been

Table 1. Student enrollment and school composition for Catholic, other private, and public schools: 1994 to 2006

	Student Enrollment				Public-to-Teacher Ratio		
	Catholic	Other Private	Public		Catholic	Other Private	Public
1994	5.73	4.72	89.55	1994	18.12	11.69	17.54
1996	5.67	4.74	89.60	1996	17.89	11.22	17.41
1998	5.58	4.87	89.56	1998	17.33	10.85	16.94
2000	5.38	4.81	89.81	2000	16.76	16.76	16.25
2002	5.26	5.13	89.61	2002	16.18	10.35	16.07
2004	4.88	4.93	90.18	2004	15.47	9.93	16.14
2006	4.56	5.07	90.37	2006	15.05	9.71	15.85
		% Hispanic				% Black	
1994	10.66	5.25	12.39	1994	8.81	9.70	16.15
1996	10.07	4.32	13.26	1996	7.91	8.20	16.45
1998	10.04	4.21	14.22	1998	7.65	8.12	16.64
2000	10.59	10.59	15.40	2000	7.82	7.82	16.82
2002	11.21	4.63	16.94	2002	7.82	8.73	16.84
2004	11.21	5.00	18.34	2004	7.60	8.73	16.88
2006	11.86	5.28	19.66	2006	7.43	8.81	16.81

Source: Common Core and Private School Universe Survey Data.

steadily losing students since 2000 and, after 2002, the loss rate did increase—by about one fifth of 1% of all students nationally. Thus, the post-2002 Catholic school enrollment loss is so small relative to public school enrollment that it could pass as sampling error in the public school data where no obvious increase is discernible after 2002. Nor did enrollment in non-Catholic private schools suddenly increase around 2002, undercutting the argument that many Catholic school students exited to enroll in other private schools. Table 1 also provides details about three variables likely to suddenly affect mean achievement changes after 2002: pupil-to-teacher ratios, percentage Hispanic, and percentage Black students. None appears to deviate from its secular trend immediately after 2002, making it hard to argue that the composition of public and Catholic schools suddenly changed after 2002 in ways that correlate with academic achievement. Even so, the second contrast we present offers an even better way of ruling out the Catholic sex scandal threat.

Contrast 2: Public versus non-Catholic private schools. Non-Catholic private schools experienced no corresponding sex scandal. Nor is there clear empirical evidence that the children leaving Catholic schools around 2002 transferred into non-Catholic private schools. Nor is there evidence in non-Catholic private schools of a sudden 2002 change in student–teacher ratios or in the numbers of Black or Latino students. So the non-Catholic private schools provide a better counterfactual than the Catholic schools.

Contrast 3: High-standard versus low-standard states. Under NCLB, states use their own achievement tests, proficiency cutoffs, and timetables for attaining total state proficiency by 2014. In the program's first years, states using difficult tests, high cutoffs, and a front-loaded timetable had more schools failing AYP that therefore had to undertake more frequent and more stringent school changes. Conversely, states choosing easier tests, lower cutoffs, and a back-loaded timetable tended to have fewer failing schools and, on

Table 2. Percentage of students proficient on State and National Assessment of Educational Progress (NAEP) Tests: 2003

High-Proficiency Standard States			Low-Proficiency Standard States		
State	State Test	NAEP Test	State	State Test	NAEP Test
Arizona	46	24	Colorado	83	35
Arkansas	46	25	Connecticut	76	39
California	36	22	Georgia	76	25
District of Columbia	48	8	Minnesota	75	40
Hawaii	31	21	Nebraska	80	33
Kentucky	47	27	New Hampshire	76	39
Maine	35	34	North Carolina	85	34
Massachusetts	50	41	Tennessee	80	24
Missouri	29	32	Texas	84	28
Rhode Island	45	28	Virginia	75	35
South Carolina	26	27	Wisconsin	77	35
Washington	46	34			
Wyoming	39	35			
M	40	27		79	33

Note. Results are averaged across grades (fourth and eighth grade), subjects (math and reading) in year 2003 for state and NAEP assessment. When state assessment data are missing in the grade examined, data from the next lower grade are used and if not available then data are from the next higher grade. *Source:* Consolidated State Performance Report and Institute of Education Science.

average, these had failed for fewer years and so had to make fewer and less profound educational changes. This reality helps describe the third contrast based on variation in proficiency standards across states.

We measure these standards by averaging each state's student proficiency rate in 2003 across two grades (fourth and eighth grade) and two subject areas (reading and math).[1] This aggregation increases reliability and provides a continuous measure of state proficiency with a passing rate that varies from 26% to 85% across states. To reduce dependence on functional form assumptions and to present results more intuitively, we also constructed a trichotomous variable. States with fewer than 50% of students meeting proficiency standards became the high standards (HS) group ($N = 13$, including the District of Columbia), states with 75% or more became the low standards (LS) group ($N = 11$), and the rest became the medium proficiency standards group ($N = 25$). So we have two nonindependent measures of state proficiency—one continuous and the trichotomous. Table 2 shows the state names.

The state-level dosage tests have some advantages over the national tests in Contrasts 1 and 2. First, any sources of bias affecting the public versus private school contrast after 2002 cannot apply in any simple way to the state-level contrast, for it only includes public schools. Second, Contrasts 1 and 2 involve at most 20 degrees of freedom—10 time points by two nonequivalent groups—and so statistical power is less than desirable. Fortunately, Contrast 3 includes 48 states assessed up to 10 times, resulting in 384 to 432 state/time observations, the exact number depending on the number of years when a given achievement

[1] We exclude New York because it uses its own state proficiency scale that is not based on a 0 to 100% proficiency scale that other states use. We also exclude Vermont because it has no state assessment data for the years examined.

outcome was assessed. Of course, the state data might be more intertemporally variable than the national data, and state variation in NCLB dosage may have less impact than NCLB's presence versus absence in the national contrasts. So greater power is not inevitable, though it does turn out to be the case in the data we report. But Contrast 3 entails a different causal question from Contrasts 1 and 2. Instead of asking about the national effects of NCLB when present versus absent, it asks about the effects of state variation in NCLB dosage.

The contrast of HS versus LS states raises some validity issues. First, state variation in proficiency standards may reflect true state achievement differences rather than a standards-setting strategy. But that is hardly the case. Table 2 shows the percentage of proficient students in HS and LS states on both NAEP and state tests. The average NAEP difference between the HS and LS states is modest (27% proficient vs. 33%). However, the difference on states' own achievement tests is very large by comparison (40% vs. 79% proficient). In addition, the small true state differences in NAEP were accounted for in some outcome models we report that used the difference between state and NAEP test scores in 2003. This made little difference to the basic pattern of results compared to when only NAEP 2003 data were used, though effects did tend to be smaller and fewer were statistically significant.

Since NCLB was passed in January 2002, a second concern is with using 2003 proficiency data to define state standards. The later program onset date is partly justified because the law did not come into operation in schools until fall 2002. Even so, standards might have changed earlier in anticipation of NCLB. Table 3 provides data on state proficiency rates from 2001 to 2005, and Table 4 shows the correlation of the percentage of students proficient by years both before and after NCLB. All of the between-year correlations exceed 0.80, though they are slightly lower around 2002 than at other times. As Table 3 shows, most states left their standards intact after 2002. But six did significantly lower them—two in the HS, two in the LS, and two in the medium states—and so not likely to entail any bias. Only Hawaii significantly increased standards, but it was in the HS group anyway. But as a further precaution, we ran all the outcome models using 2001 to define state standards as well as 2003. Given the high correlation between years, the earlier date made little difference. Although NCLB passed in 2002, 2003 is close to bias-free for defining state standards.

Classifying states as HS or LS makes sense only if HS states actually undertook more fundamental educational reforms. Data to test this come from Consolidated State Performance Reports, but only beginning in 2007. Data from that year are in Table 5, and analyses with schools as the unit of analysis show that HS states have indeed undertaken more reform. By that year, most schools in most states were not failing AYP. However, 13% more schools in HS states are classified as "in need of improvement" for failing to make AYP in at least 2 consecutive years. Sixteen percent more students are eligible for the school choice provision and 8% more for supplemental services. Of schools that failed AYP, more are taking on corrective and restructuring actions in HS states—4% more have instituted a new curriculum and 9% more have taken alternative types of restructuring. Schools are free to choose more than one reform activity. If all schools undertaking corrective or restructuring items took a single action, then the difference between HS and LS states would be 9% in both. If schools made all the suggested reforms simultaneously, then the difference is 5% in corrective action and 8% in restructuring. In 2007, HS states definitely have more failing schools and undertake more reforms engaging fundamental aspects of school life.

Three Outcomes. For each of these three national or state-level contrasts we assess how stable causal results are across fourth- and eighth-grade math achievement and fourth-grade

Table 3. Student proficiency rates by state: 2001–2005

State	Group	SA01	SA02	SA03	SA04	SA05
Alabama	Med	67	.	61	69	72
Alaska	Med	77	67	70	70	73
Arizona	High	48	.	46	48	64
Arkansas	High	37	.	46	54	48
California	High	33	31	36	37	42
Colorado	**Low**	**54**	**55**	**83**	**84**	**85**
Connecticut	Low	76	76	76	76	74
Delaware	Med	65	68	68	70	72
District of Columbia	**High**	**24**	**21**	**48**	**46**	**45**
Florida	**Med**	**58**	**21**	**56**	**59**	**60**
Georgia	Low	69	72	76	78	78
Hawaii	**High**	.	**58**	**31**	**33**	**35**
Idaho	Med	.	.	70	79	83
Illinois	Med	63	64	62	66	67
Indiana	Med	69	68	70	70	72
Iowa	Med	71	71	73	74	77
Kansas	Med	63	63	69	72	76
Kentucky	High	43	45	47	52	53
Louisiana	Med	53	49	57	58	60
Maine	High	34	34	35	35	41
Maryland	**Med**	**38**	**31**	**56**	**63**	**68**
Massachusetts	High	47	48	50	52	49
Michigan	Med	63	57	58	62	68
Minnesota	**Low**	**51**	**49**	**75**	**70**	**76**
Mississippi	Med	58	62	67	73	70
Missouri	High	29	30	29	30	32
Montana	Med	74	70	73	58	64
Nebraska	Low	75	68	80	84	88
Nevada	Med	53	51	53	47	49
New Hampshire	**Low**	**33**	**34**	**76**	**76**	.
New Jersey	Med	70	70	69	72	74
New Mexico	Med	38	.	67	52	42
North Carolina	Low	81	84	85	87	88
North Dakota	Med	74	60	61	66	73
Ohio	Med	59	61	61	67	69
Oklahoma	Med	66	65	67	70	71
Oregon	Med	69	71	70	70	75
Pennsylvania	Med	55	55	57	63	65
Rhode Island	High	55	.	45	52	.
South Carolina	High	26	29	26	31	32
South Dakota	Med	57	57	72	77	80
Tennessee	Low	.	.	80	82	87
Texas	Low	92	93	84	81	76
Utah	Med	64	68	74	75	76
Virginia	Low	71	73	75	78	81
Washington	High	44	48	46	60	65
West Virginia	Med	58	.	65	75	77
Wisconsin	Low	64	67	77	75	77
Wyoming	High	.	.	39	44	41

Note: Bolded text highlight states where proficiency rates increased by over 20 percentage points.
Source: U.S. Department of Education.

Table 4. Correlation of state assessment proficiency rates: 2001 to 2005

	Year 2001	Year 2002	Year 2003	Year 2004
Year 2001				
Year 2002	0.91			
Year 2003	0.85	0.82		
Year 2004	0.81	0.79	0.95	
Year 2005	0.80	0.77	0.94	0.98

reading. This replication is important for external validity purposes, for possibly demarcating causal boundary conditions that a highly standardized study with a single grade and subject matter cannot probe (Fisher, 1935). But the three grades and subject matters are also important for statistical power. The public versus private school contrasts entail so few degrees of freedom that the three contrasts and three outcomes enable us to test for coherence across the direction of causal coefficients as well as for the statistical significance of results. And because unique samples of schools were involved in collecting the data across both grades and achievement domains, the opportunity arises to analyze many independent tests of the direction of causal influence.

Main and Some Trend NAEP Time Series. None of the design elements just added speaks to time series. However, we use two independent national data sets to assess differences of differences in mean and slope across the various contrasts, grades, and achievement domains just listed. One data source is Trend NAEP, where the items do not change over time; the other is Main NAEP, where items do sometimes change to reflect shifts in national curricula.

Using Main NAEP would be problematic if its testing regimen suddenly changed after 2002. The item selection did not change then. However, the sampling design did in order to reduce the number of schools participating. Fortunately, this was achieved by randomly drawing schools from those already randomly selected, and so no bias is introduced (Lazer, 2004; National Center for Education Statistics, 2009). But for changes in sampling design between 2002 and 2004 to explain NCLB effects requires that these changes occurred not just differentially between public two types of private schools but also between states whose implementation of NCLB led to more or less fundamental school reform. To rule out these remote possibilities, we also used Trend NAEP whose items remain constant.

Trend NAEP is more limited in its usage than Main NAEP because its design precludes state-level estimates, because national data are not available for non-Catholic private schools during the postintervention years, and because accommodations for students with special needs began in 2004. Although data using both the old and the new sampling frames were collected in that year, this was not repeated. As a result, 2004 is the last data point with the same population as the pre-NCLB time series. Even so, Trend NAEP still allows replication of any immediate mean changes observable in Main NAEP for the public versus Catholic school contrast. This adds an empirical rationale to the prior logical ones for ruling out the effects of changes in testing procedures.

Testing Three Types of Causal Hypothesis. For each contrast we test three kinds of longitudinal DD hypothesis. Using formal statistical models presented later, we evaluate the following:

Table 5. Percentage of schools and teachers meeting NCLB requirements in 2007: By level of state proficiency standards

	High	Med	Low	H vs. L Diff
AYP Determination				
Schools that Failed to Make AYP	35.88	27.44	21.25	14.63*
Title I Schools that Failed to Make AYP	37.15	26.78	21.12	16.03*
Schools that were Unsuccessful in Appealing their Failed AYP Status	70.43	50.54	61.95	8.47*
Teacher Certification				
Deemed Not Highly Qualifed Teachers	5.95	6.93	2.41	3.54*
Schools in Need of Improvement Status (Based on All Schools)				
Schools in Early Improvement Status	10.43	8.58	4.97	5.46*
Schools in Corrective Action	2.11	1.45	1.02	1.09*
Schools in Restructuring	2.16	2.12	0.71	1.45*
Schools in Need of Improvement	14.69	12.15	6.70	8.00*
Schools in Need of Improvement Status (Based on Title I Schools Only)				
Schools in Early Improvement Status	18.05	15.39	8.91	9.14*
Schools in Corrective Action	3.65	2.60	1.82	1.82*
Schools in Restructuring	3.73	3.79	1.27	2.46*
Schools in Need of Improvement	25.43	21.78	12.00	13.43*
Based on All Title I Schools that Did Not Make AYP				
In Need of Improvement Status in Year 1 and 2: Initial Actions				
Students Eligible for School Choice	21.43	10.19	5.04	16.39*
Students Eligible for Supplemental Services	11.40	7.13	2.92	8.49*
In Need of Improvement Status in Year 3: Corrective Actions				
Institute New Curriculum (CA)	7.16	5.06	2.76	4.40*
Appointed Outside Expert Advice (CA)	5.99	2.54	4.63	1.36*
Decreased Management Authority (CA)	3.12	1.75	0.78	2.33*
Replaced School Staff (CA)	1.70	1.38	1.19	0.51*
Extended School Day (CA)	1.44	1.64	0.75	0.69*
Restructure their Internal Organization (CA)	3.22	2.43	2.72	0.49*
Replaced Principal (CA)	0.04	2.71	0.48	−0.44*
Total Corrective Actions (assuming no overlap)	22.66	17.51	13.32	9.34
Total Corrective Actions (assuming full overlap)	8.75	4.41	4.23	4.52
In Need of Improvement Status in Year 4: Restructuring Actions				
Taken Over By the State (RA)	0.99	0.63	0.00	0.99*
Replaced School Staff (RA)	1.13	1.59	0.19	0.95*
Contract Private Company to Run School (RA)	0.71	0.56	0.00	0.71*
Reopen as a Charter Schools (RA)	0.06	0.04	0.11	−0.05
Took Other RA Actions (RA)	7.79	4.44	1.72	6.07*
Total Restructuring Actions (assuming no overlap)	10.68	7.26	2.01	8.67
Total Restructuring Actions (assuming full overlap)	9.80	3.84	1.69	8.11

*Hypothesis test is conducted at the school level between high and low performance standard states and is statistically significant at $\alpha = .05$.

Source: 2007 Consolidated State Performance Report.

1. Whether in the first year after 2002, obtained public school achievement *means* differ from what pretest means and trends predict them to be by more than the same difference observed in untreated Catholic and non-Catholic private schools, and whether these same mean DD effects are obtained in HS versus LS states.
2. Whether public school achievement *time trends* after 2002 differ from their pretreatment trends by more than the difference found in Catholic and non-Catholic private schools, and whether this same DD in slopes is observed in HS versus LS states.
3. Whether, *at the last available post-NCLB time point*, the difference between predicted and obtained group means in public schools is greater than the same difference in Catholic and non-Catholic private schools and is greater in HS than LS states.

Hypothesis 3 combines the DD tests of both means and slopes in Hypotheses 1 and 2. So it involves a presumptively larger treatment contrast, but one that is not independent of the unique mean and slope DD tests.

Summarizing Study Purposes

We seek to estimate the coherence among results involving 39 tests of the effects of NCLB. Using Trend NAEP there are three tests of immediate mean effects by grade and achievement domain. Using Main NAEP, there are nine tests of public versus Catholic schools involving three grades and achievement domains and DDs in means, slopes, and the two combined. Another nine tests emerge from contrasting public and non-Catholic private schools over hypothesis types, grades, and domains. Nine more tests come from comparing HL and LS states using the trichotomous measure of state standards, and the final nine come from substituting the continuous measure of state standards. Of course, assessing the empirical coherence of results requires placing more weight on *independent* tests within this set of 39; we do that later.

Even if a high level of empirical coherence were achieved, its interpretation is still dependent on theoretical coherence, on a set of hypotheses that no other theory can predict even if it does predict some hypotheses from within the set. We later discuss three such possible alternatives: (a) a 2002 increase in the rate of students exiting Catholic schools, (b) school officials deliberately manipulating NAEP scores after 2002, and (c) new national math standards that were released in 2000. Assuming empirical coherence, the challenge these contending theories face is to explain why achievement should change in public schools after 2002 more than it changed in both Catholic and non-Catholic private schools and why this same DD should be observed at the same time in HS relative to LS states.

METHODS

Data

Most of the outcome data come from Main NAEP. It uses stratified random sampling to obtain a representative sample of students at the state and national level. It began in 1990 and we continue it until 2011 for public and Catholic schools. But because of data availability we continue it only through 2007 for non-Catholic schools in reading and through 2009 in math. Main NAEP seeks to represent the nation's current instructional practices, and so test content changes over time. No significant item changes were made in 2002, but the sampling design was modified then to reduce the number of participating schools. But, as noted earlier, this was done without compromising the crucial random sampling feature.

Main NAEP requires a participation rate of at least 70% for its results to be considered reliable. This has not been a problem with public or Catholic schools, but in some years it has been with non-Catholic private schools. Nonetheless we use estimates for the less reliable years, and they may be responsible for the somewhat greater variability observed around the non-Catholic private school time trends. Yet, as readers see, impact estimates are largely similar when public schools are contrasted with Catholic and with non-Catholic private schools.

Content changes in Main NAEP were made in 1996 and 2000 for math and in 1998 for reading to accommodate students with disabilities and special English language needs. To assess this population change, the test developers used a split sample design in the change year, one sample having the new accommodations and the other not. We analyzed the achievement data both ways and ascertained there was no difference. The analyses we present use the pre-NCLB data with accommodations, as this is the same as the post-NCLB population.

Trend NAEP has a different purpose from Main NAEP. It aims to assess national performance over time and so does not change its test content, an obvious advantage in ITS work. However, Trend NAEP is less useful than Main NAEP because, as noted, it is not collected at the state level, is not publicly available for non-Catholic private schools in the post-NCLB years, and the intervals between observations are generally longer. Moreover, accommodations for students with special needs began in 2004 when data were collected using both the old and new sampling frames. However, they were not collected together subsequently, and so 2004 is the last data point with the same population as the pre-NCLB time series. So we cannot use Trend NAEP post-2004. This means we can replicate any immediate mean differences found with Main NAEP in the public versus Catholic schools contrast but cannot examine differences in slope differences.

Analytic Models

A linear regression model was used to analyze the public versus private school contrasts and a fixed effects model to analyze the state contrasts. Each model includes a dummy variable for the relevant contrast and interactions of this dummy with the change in average test score and growth rate from before to after NCLB. Thus, each analysis has one time-series segment representing the mean and growth in achievement scores prior to NCLB and the other representing the corresponding means and slopes after 2002.

We selected 2002 because most school districts did not begin implementing NCLB until fall of that year. NAEP tests are conducted in January through March of each testing year, and so we treat the 2002 testing (available only for reading) as preceding the school-level implementation of NCLB. For math, the first postintervention data point is 2003, clearly a post-NCLB year.

For the contrast of public schools with either Catholic or non-Catholic private schools we estimate the following regression model:

$$\begin{aligned} Y_{tj} = {} & \beta_0 + \beta_1(year)_{tj} + \beta_2(group)_j + \beta_3(policy)_{tj} + \beta_4(year \times group)_{tj} \\ & + \beta_5(policy \times year)_{tj} + \beta_6(policy \times group)_{tj} \\ & + \beta_7(policy \times year \times group)_{tj} + \varepsilon_{tj}, \end{aligned} \tag{1}$$

where Y_{tj} is the outcome on Main NAEP at $t = 1,\ldots,10$ time points for fourth-grade reading, for example, and Trend NAEP at $t = 1,\ldots,6$ time points for fourth-grade reading, for example. *Year* is a continuous variable indicating the year of measurement; *group* is a dichotomous variable representing public ($group = 1$) versus private schools ($group = 0$); $policy$ is another dichotomous variable indicating pre- and post-NCLB period ($1 =$ post-NCLB). Hypothesis tests of group differences in mean and slope changes require including all two- and three-way interactions, and so the regression coefficient β_6 of the $policy \times group$ interaction gives the differences in the mean change in 2002. The three-way interaction effect β_7 tests whether public and private schools differ in their post-NCLB slope changes. The error term ε_{tj} is assumed to be independent and identically distributed according to a normal distribution with a mean of zero and variance $\sigma\frac{2}{\varepsilon}$. No covariates are included, given the limited degrees of freedom for the public versus private contrast.

For the categorical contrast of states with high, medium, and low proficiency standards, we estimate a model with state and year fixed effects. The outcomes are modeled as

$$\begin{aligned} Y_{ti} = {} & \beta_0 + \beta_1\,(year \times group\_h)_{ti} \\ & + \beta_2\,(year \times group\_m)_{ti} + \beta_3\,(policy \times group\_h)_{ti} \\ & + \beta_4(policy \times group\_m)_{ti} + \beta_5(policy \times year \times group\_h)_{ti} \\ & + \beta_6(policy \times year \times group\_m)_{ti} + \beta_7(percent\_free\_lunch)_{ti} \\ & + \beta_8(pupil\_teacher\_ratio)_{ti} + \mu_{i\_} + \tau_{ti} + \varepsilon_{ti} \end{aligned}$$

where *group_h* and *group_m* in the interaction terms are dummy variables indicating high and medium performance standard states, respectively. μ_i are the state fixed effects, τ_t the year fixed effects that allow for a flexible modeling of the nonlinear functional form, and $\varepsilon_{tj} \sim N(0, \sigma_\varepsilon^2)$ the independent and identically distributed error term.[2] Again, hypothesis tests of differences in mean and slope changes between groups are represented by two- and three-way interactions. The regression coefficient β_3 of the $policy \times group\_h$ interaction gives the differences in the mean change in 2002 for the high vs. low proficiency standards group. The three-way interaction effect β_5 tests whether high and low proficiency standards group differ in their post-NCLB slope changes. The model also controls for time-varying covariates assessing the percentage of students eligible for free lunch and the pupil-to-teacher ratio (*percent_free_lunch, pupil_teacher_ratio*).

We examined other potential time-varying covariates at both the state and national level, including school-level expenditures, family income and various student demographics. They all correlate quite highly with the percentage of students eligible for free lunch or pupil-to-teacher ratio and do not change treatment effects when included. Therefore, the final model is restricted to the latter two covariates that are largely independent of each other yet are highly correlated with the achievement outcomes.[3]

To take serial correlation into account, we computed robust standard errors using the CLUSTER option in STATA (Rogers, 1993). We also did ARIMA modeling despite the limited number of temporal data points and found the best approximation to be a second order autoregressive model. Because the outcome results we present hardly differed by

[2] We also ran all analyses using a random effects model and got essentially the same results with respect to both point estimates and statistical significance levels.

[3] The models were not weighted by student population or the inverse sampling variance of NAEP estimates. This was because states are the unit of analysis and little variation results when standard deviations are examined separately across states and years.

the type of correction for serial autocorrelation, we report here the results with clustered standard errors.

Null Hypotheses

For all models, the null hypotheses are as follows:

Hypothesis 1: Group mean differences do not differ from before to after NCLB.

H1:$\beta_6 = 0$ for the public versus either private school contrast (Equation 1)

H1:$\beta_3 = 0$ for the high versus low proficiency standard states contrast (Equation 2)

Hypothesis 2: Group slope differences do not differ from before to after NCLB.

H2:$\beta_7 = 0$ for the public versus either private school contrast (Equation 1)

H2:$\beta_5 = 0$ for the high versus low proficiency standard states contrast (Equation 2)

Hypothesis 3: Group mean differences at the final endpoint available (2007, 2009, or 2011 depending on the measure and contrast) do not differ from what the pre-2002 group mean and slope trends predict them to be. This total impact hypothesis combines the first two hypotheses about differences in mean and slope.

H3: $\beta_6 + (\beta_7\ k) = 0$ for the public versus either private school contrast

H3: $\beta_3 + (\beta_5\ k) = 0$ for the high versus low proficiency standard states contrast

with $k = 5$, 7 or 9 for Main NAEP depending on the contrast and outcome of interest.

RESULTS

Reading and math results are presented as raw achievement scores and are also transformed into effect sizes,[4] percentile rank gains,[5] and months of learning.[6] In the figures, we present raw means and best fitting regression lines.

Trend NAEP: Fourth- and Eighth-Grade Math Results

Figure 1 shows the fourth-grade math effects for *Trend NAEP*, comparing public and Catholic schools. Pre-NCLB, there are two essentially parallel trend lines, with the Catholic school mean being higher. After NCLB, the two groups no longer differ, and the gap between them is eliminated. The first row of Table 6 shows that the post-NCLB mean difference is significantly different from the extrapolation of the pre-NCLB trend difference.

Figure 2 provides the eighth-grade math results. The Catholic school mean is again higher initially, though the two groups are now growing further apart. The post-NCLB

[4] We use NAEP-provided grade- and subject-specific standard deviations from individual student test score data.

[5] Percentile rank gains reflect the number of ranks a state would have risen relative to other states by virtue of its NCLB gains. Gains in percentile rank are based on the distribution of state rank in 2002.

[6] We translate the study's obtained effect sizes to months of learning. Analyses of nationally normed tests by Hill et al. (2007) show that the average annual test score gain in effect size from fourth to fifth grade is roughly 0.40 standard deviation units for reading and 0.56 for math. A much smaller gain of 0.22 is observed for the average test score gain from eighth-grade to ninth-grade math. So an obtained effect size of 0.20 *SD* in fourth-grade reading translates to 6 months' worth of learning based on the benchmark effect size of 0.40 (i.e., 0.20/0.40 × 12 months). But the same effect size will translate into many more months of learning in eighth grade because of smaller benchmark effect sizes.

Table 6. NCLB National effects on means, slopes and by final endpoint for Contrasts 1, 2 and 3 with Main National Assessment of Educational Progress (NAEP) and for Contrast 1 with Trend NAEP

	DD Model Effects			Effect Size		
	M	SD	t Value	SD[a]	Months[b]	%[c]
Fourth-grade math						
Public vs. Catholic (Trend NAEP)[d]						
Diff. in mean difference (2004)	10.93	5.53	1.97*	0.34	7.20	—
Public vs. Catholic (Main NAEP)						
Diff. in mean difference	3.93	3.82	1.03	0.15	3.18	0.15
Diff. in slope difference	1.00	0.55	1.82	0.04	0.81	0.09
Diff. by final endpoint	12.95	6.27	2.07*	0.49	10.47	0.57
Public vs. non-Catholic (Main NAEP)						
Diff. in mean difference	4.41	6.11	0.72	0.17	3.56	0.17
Diff. in slope difference	0.29	1.03	0.28	0.01	0.24	0.04
Diff. by final endpoint[e]	6.46	8.39	0.77	0.24	5.22	0.38
Eighth-grade math						
Public vs. Catholic (Trend NAEP)						
Diff. in mean difference (2004)	7.26	2.03	3.58*	0.24	12.99	—
Public vs. Catholic (Main NAEP)						
Diff. in mean difference	1.54	4.15	0.37	0.05	2.47	0.04
Diff. in slope difference	0.30	0.60	0.50	0.01	0.47	0.01
Diff. by final endpoint	4.20	6.81	0.62	0.12	6.73	0.17
Public vs. non-Catholic (Main NAEP)						
Diff. in mean difference	0.77	5.79	0.13	0.02	1.24	0.02
Diff. in slope difference	1.48	0.98	1.51	0.04	2.38	0.04
Diff by final endpoint	11.16	7.95	1.40	0.33	17.90	0.49
Fourth-grade reading						
Public vs. Catholic (Trend NAEP)						
Diff. in mean difference (2004)	3.92	3.28	1.20	0.11	3.22	—
Public vs. Catholic (Main NAEP)						
Diff. in mean difference	0.77	3.02	0.25	0.02	0.67	0.03
Diff. in slope difference	0.57	0.51	1.12	0.02	0.50	0.03
Diff. by final endpoint	5.93	4.88	1.21	0.17	5.16	0.41
Public vs. non-Catholic (Main NAEP)						
Diff. in mean difference	0.95	4.25	0.22	0.03	0.83	0.04
Diff. in slope difference	0.22	1.10	0.20	0.01	0.19	0.01
Diff. by final endpoint	2.06	4.68	0.44	0.06	1.79	0.08

Note. DD = difference-in-differences design.

[a]Effects sizes are computed using group averaged grade- and subject-specific standard deviations of student test scores provided by NAEP. For Main NAEP Catholic vs. public analyses, *SD* = 35 for fourth-grade reading, *SD* = 27 for fourth-grade math, *SD* = 34 for eighth-grade math. For Main NAEP other private vs. public analyses, *SD* = 35 for fourth-grade reading, *SD* = 27 for fourth-grade math, *SD* = 35 for eighth-grade math. For Trend NAEP Catholic vs. public analyses, *SD* = 37 for fourth-grade reading, *SD* = 33 for fourth-grade math, *SD* = 31 for eighth-grade math. [b]Gains in months are based on the average grade- and subject-specific effect size in moving from one grade to next on nationally normed tests. [c]Gains in percentile rank are calculated based on the distribution of state ranking observed in 2002. [d]Percentile cannot be calculated because no distributional data on Trend NAEP are available at the state level. [e]Reading estimates for 2007, math estimates for 2009.

$^{*}p < .05$.

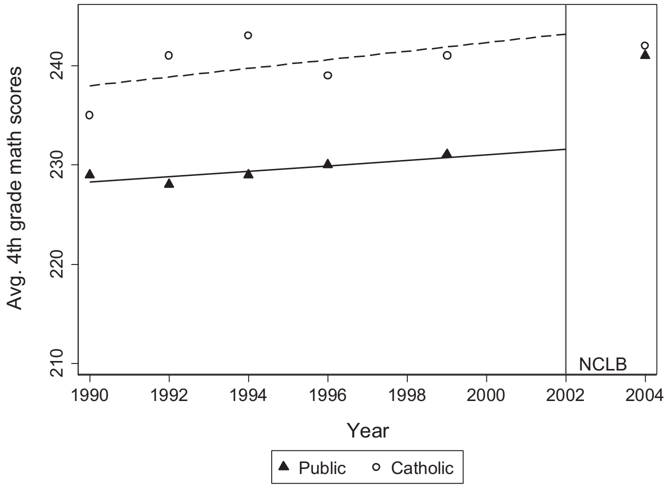

Figure 1. Fourth-grade math scores for Trend National Assessment of Educational Progress: Public and Catholic schools. *Note.* NCLB = No Child Left Behind.

difference is again less than earlier, indicating that public schools narrowed the gap after 2002. This difference was again statistically significant (second row of Table 6).

Main NAEP: Fourth- and Eighth-Grade Math Results

The *Main* NAEP results for public versus Catholic schools are in Figure 3 for fourth graders and Figure 4 for eighth graders. Both show a growing gap prior to NCLB and a reduction afterward, although it is much clearer for the fourth than the eighth grades. All coefficients are positive, though only the fourth-grade total effect is statistically significant (see Table 6).

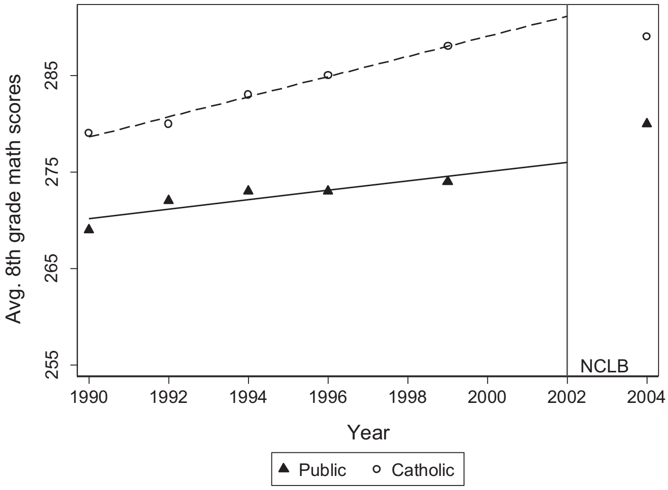

Figure 2. Eighth-grade math scores for Trend National Assessment of Educational Progress: Public and Catholic schools. *Note.* NCLB = No Child Left Behind.

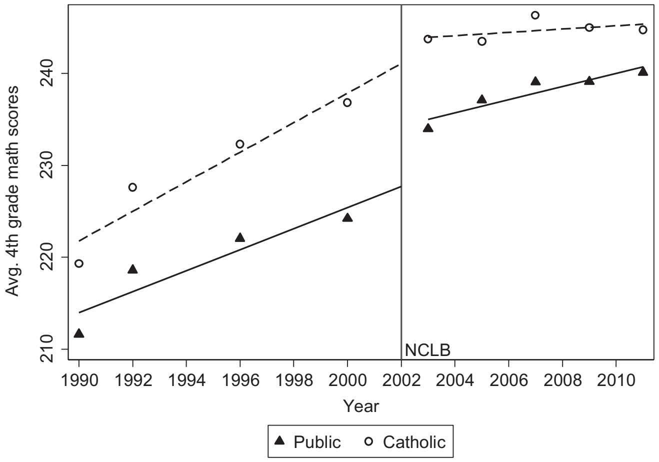

Figure 3. Fourth-grade math scores for Main National Assessment of Educational Progress: Public and Catholic schools. *Note.* NCLB = No Child Left Behind.

Nonetheless, the pattern of mean and slope differences in differences in math is similar across grades, and the direction of immediate math mean DDs is similar in direction for both Main and Trend NAEP.

The Main NAEP math results for the public versus non-Catholic schools are in Figure 5 for fourth graders and Figure 6 for eighth graders. Again we see that public schools are initially behind but reduce the math gap. Once again, all coefficients for the shift in mean, slope, and total change are positive. But, as Table 6 shows, no immediate or slope effects are statistically significant. However, the total change is significant for fourth-grade math, as is the total change for eighth-grade math when the somewhat deviant 1990 value is omitted.

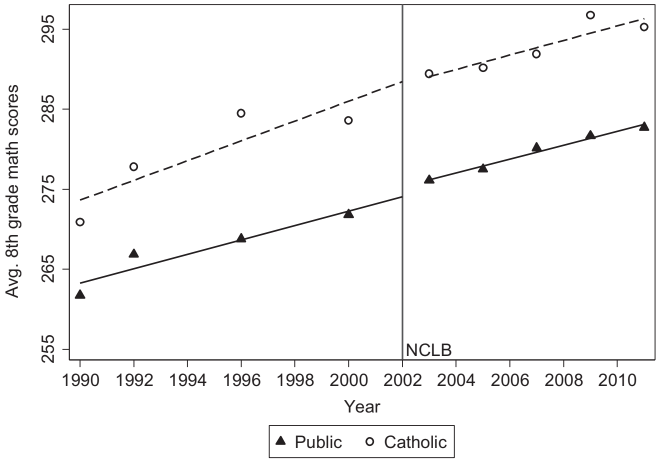

Figure 4. Eighth-grade math scores for Main National Assessment of Educational Progress: Public and Catholic schools. *Note.* NCLB = No Child Left Behind.

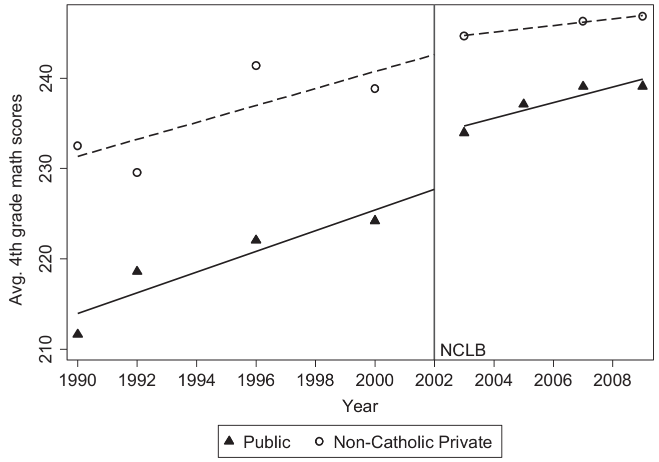

Figure 5. Fourth-grade math scores for Main National Assessment of Educational Progress: Public and non-Catholic private schools. *Note.* NCLB = No Child Left Behind.

Turning to Contrast 3 and the within-state results, Figure 7 gives the fourth-grade results and Figure 8 the eighth-grade math ones. The consistent pre-NCLB picture is of HS schools performing worse initially and changing more slowly over time. However, this pattern is reversed after NCLB, and the growing gap is again narrowed. The differences in both mean and slope differences are evident to the eye and are corroborated in the statistical tests in Table 7. All coefficients are in the expected direction; for fourth-grade math, the mean and total change estimates significantly differ from zero; and in eighth grade the slopes and total change estimates also do so. The size of these effects is also in Table 7.

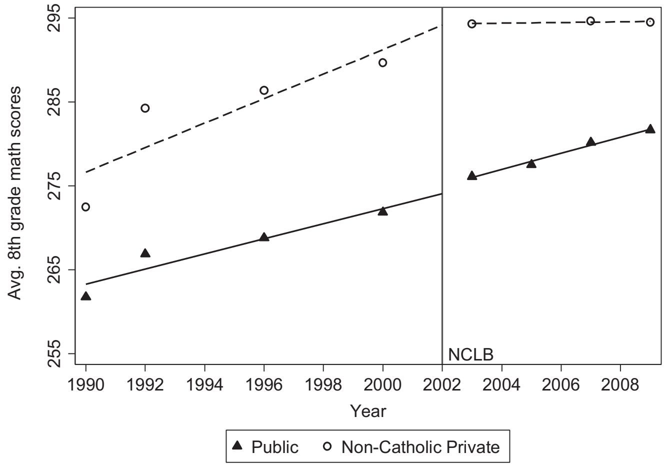

Figure 6. Eighth-grade math scores for Main National Assessment of Educational Progress: Public and non-Catholic private schools. *Note.* NCLB = No Child Left Behind.

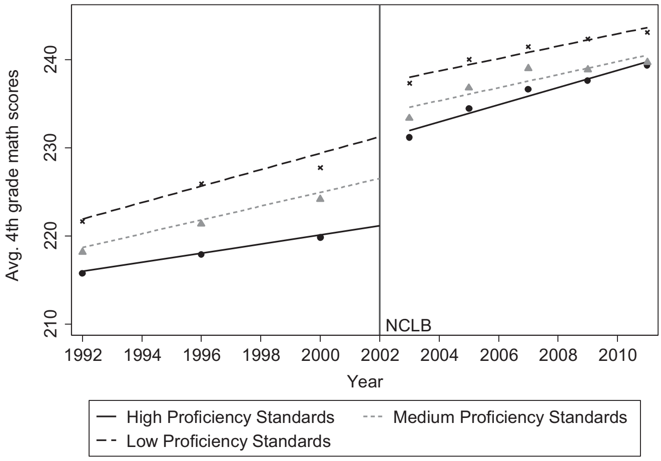

Figure 7. Fourth-grade math scores for Main National Assessment of Educational Progress: High versus medium versus low proficiency standard states. *Note.* NCLB = No Child Left Behind.

To summarize for math: Across three contrasts, two grade levels, and three types of causal hypothesis, the math data are totally coherent with NCLB being effective. Visual inspection, the direction of coefficients, and statistical significance patterns when mean and slope changes are combined, all indicate that NCLB raised achievement and narrowed the prior math gaps favoring private schools over public ones and favoring states with lower proficiency standards over higher ones. Averaged across grades, in the contrasts involving Catholic schools the results correspond to about 8 months of total public school math gain after 2002. In the contrasts with non-Catholic private schools, the total gain is 11 months. In the state-level contrasts, the respective gains were 9 months and 8 months. In student-level standard deviation units, the average math effect is about 30 *SD*s across both grades and the state and national levels of analysis.

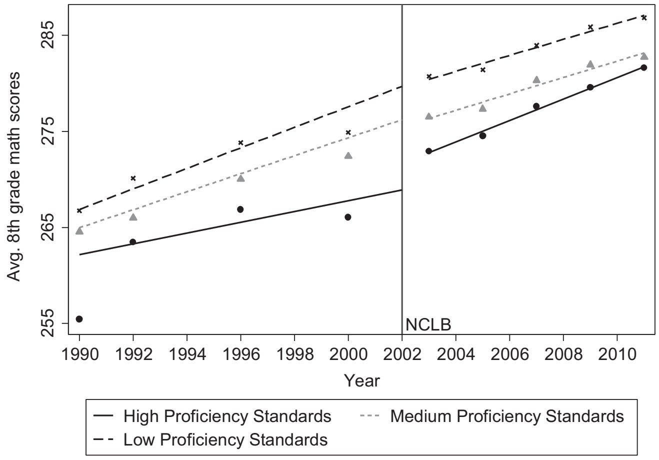

Figure 8. Eighth-grade math scores for Main National Assessment of Educational Progress: High versus medium versus low proficiency standard states. *Note.* NCLB = No Child Left Behind.

Table 7. No Child Left Behind state effects on means, slopes and by final endpoint for Contrasts 1, 2 and 3 in Main National Assessment of Educational Progress (NAEP) using a categorical or continuous state proficiency measure

	DD Model Effects			Effect Sizes		
	Coeff.	SE	t	*SD*[a]	Months[b]	%[c]
Fourth-grade math						
Categorical measure						
Diff. in mean difference	3.72	1.68	2.22*	0.13	2.85	0.19
Diff. in slope difference	0.13	0.33	0.38	0.00	0.10	0.00
Diff. by final endpoint	8.83	3.82	2.31*	0.32	6.76	0.48
Continuous Measure						
Diff. in mean difference	3.18	1.55	2.06*	0.09	1.84	0.18
Diff. in slope difference	0.45	0.33	1.36	0.01	0.26	0.00
Diff. by final endpoint	7.21	3.17	2.27*	0.19	4.18	0.31
Eighth-grade math						
Categorical measure						
Diff. in mean difference	2.49	1.98	1.26	0.07	3.77	0.09
Diff. in slope difference	0.65	0.28	2.34*	0.02	0.98	0.01
Diff. by final endpoint	8.29	3.67	2.26*	0.23	12.57	0.30
Continuous measure						
Diff. in mean difference	3.02	2.20	1.37	0.08	4.45	0.11
Diff. in slope difference	0.55	0.25	2.21*	0.01	0.81	0.01
Diff. by final endpoint	7.96	3.26	2.44*	0.22	11.73	0.29
Fourth-grade reading						
Categorical measure						
Diff. in mean difference	1.38	0.91	1.52	0.04	1.12	0.02
Diff. in slope difference	0.30	0.22	1.35	0.01	0.24	0.01
Diff. by final endpoint	4.09	2.29	1.79†	0.11	3.32	0.13
Continuous measure						
Diff. in mean difference	1.21	0.70	1.74†	0.03	0.98	0.02
Diff. in slope difference	0.26	0.20	1.32	0.01	0.21	0.01
Diff. by final endpoint	3.59	1.83	1.96†	0.10	2.91	0.08

Note. DD = difference-in-differences design.
[a] Effects sizes are computed using grade- and subject-specific standard deviations of individual student test score data provided by NAEP. *SD* = 37 for fourth-grade reading, *SD* = 28 for fourth-grade math, *SD* = 36 for eighth-grade math. [b] Gains in months are based on the average grade- and subject-specific effect size in moving from one grade to next on nationally normed tests. [c] Gains in percentile rank are calculated based on the distribution of state ranking observed in 2002.
†$p < .1$. *$p < .05$.

Fourth-Grade Reading Results

Figures 9 through 12 present the reading results for both Trend and Main NAEP. Again, all show an initial difference favoring the private or LS comparison schools over the public or HS schools. And once again, all the differences of differences in mean, slope, and endpoint indicate that reading gaps narrowed after NCLB, just as math gaps did. But the pattern of statistical significance is much weaker for reading. No national contrast is significant, and in the state contrasts only the two nonindependent end-point gains are even marginally

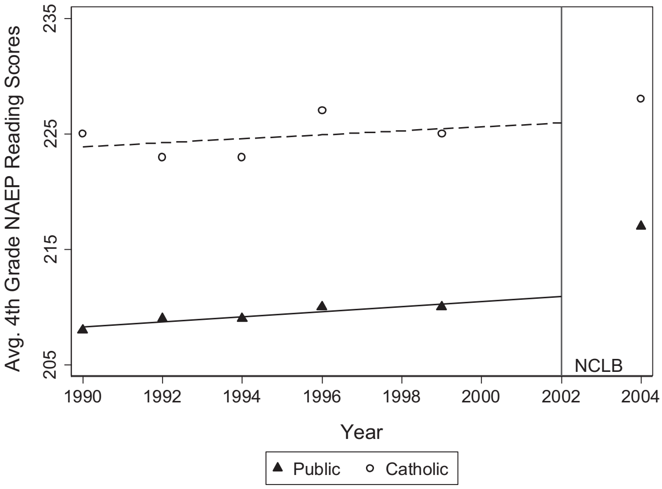

Figure 9. Fourth-grade reading scores for Trend National Assessment of Educational Progress (NAEP): Public and Catholic schools. *Note.* NCLB = No Child Left Behind.

significant. So the case for a reading effect is based on results with a consistent causal direction but weak statistical support and modest effect sizes over a decade—about 3 months of learning gain or about .10 in student standard deviation units. The math effects are consistently larger.

Coherence of the Obtained Data Pattern

The figures we have presented show a high degree of pretest consistency. Regression lines fit the data very well and functional forms are basically linear, the sole possible exception

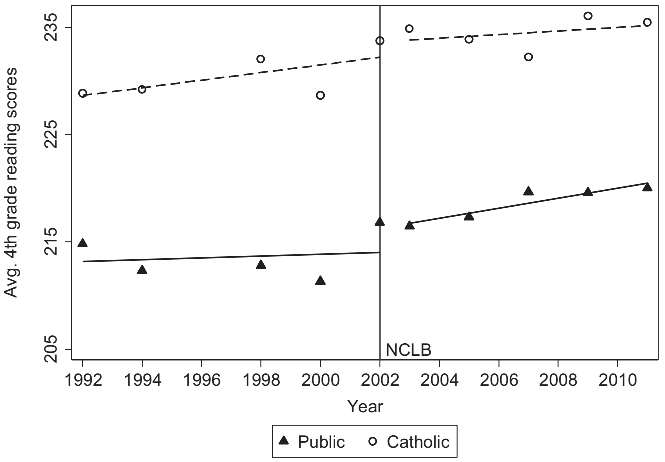

Figure 10. Fourth-grade reading scores for Main National Assessment of Educational Progress: Public and Catholic schools. *Note.* NCLB = No Child Left Behind.

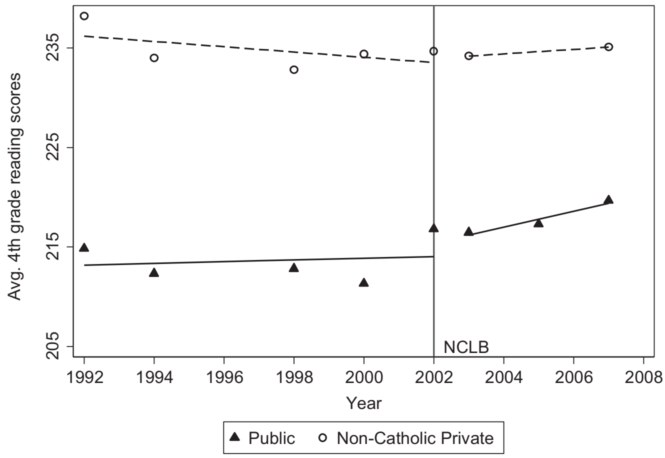

Figure 11. Fourth-grade reading scores for Main National Assessment of Educational Progress: Public and non-Catholic private schools. *Note.* NCLB = No Child Left Behind.

being the 1990 Main NAEP math values for fourth and eighth grade. The preintervention means and slopes are always lower for public than private schools and for HS than LS states, entailing a "fan spread" model of selection-maturation in each case (Campbell & Erlebacher, 1970) and entailing that a successful NCLB will probably have to narrow these growing achievement gaps. Also noteworthy is that test scores do not suddenly change just before 2002, thus ruling out statistical regression as a threat to internal validity.

The data patterns after 2002 are also coherent. All the figures and coefficients for math and reading show that achievement gaps were reduced after NCLB. Positive coefficients favoring NCLB were obtained in every one of the 39 DD tests of means (H1), slopes (H2) and the sum of the two (H3), irrespective of whether fourth or eighth grade was involved,

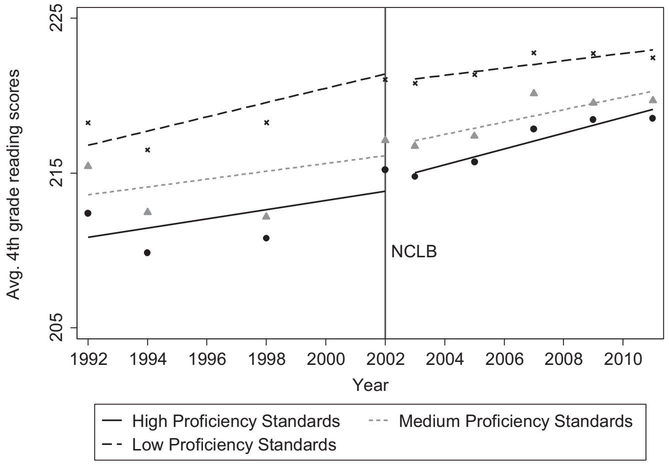

Figure 12. Fourth-grade reading scores on Main National Assessment of Educational Progress: High versus medium versus low proficiency standard states. *Note.* NCLB = No Child Left Behind.

math or English, Main NAEP or Trend NAEP, Catholic or non-Catholic private schools, and whether state proficiency differences were indexed in categorical or continuous form.

Of course, these 39 tests are not independent, precluding any simplistic use of a sign test. The dependencies include the following: Contrasts 1 and 2 share the same treatment group, hypothesis H3 is the sum of Hypotheses 1 and 2, and both state proficiency indices are constructed from the same data. Nonetheless, some subsets of hypotheses are independent. Thus, separate school samples are used for fourth- and eighth-grade NAEP math data collection, for the Trend and Main NAEP assessments, and for the state versus national NAEP estimates. The tests of mean and slope DDs are also independent.

To assess the coherence of independent math effects, in Main NAEP we have tests of immediate and slope DDs across two grades for both Contrasts 1 and 3, whereas with Trend NAEP we can test immediate effects across the two grades for Contrast 1 only. Ten independent tests result, therefore, and the results indicate that each has the same positive causal sign. The probability of this is smaller than .001. If we limit ourselves to Contrasts 2 and 3, eight independent tests of a math effect are possible—no Trend NAEP results exist for non-Catholic schools. All eight tests result in coefficients in the same positive direction ($p < .01$).

For reading, there are 15 independent tests of Contrasts 1 and 3 and 12 of Contrasts 2 and 3. All have the same positive causal sign, and this is unlikely to be due to chance.

The statistical significance levels from the tests of differences in mean and slope differences are weak. None is statistically significant when Main NAEP is used, and only the immediate math effects are for Trend NAEP. This probably reflects the modest degrees of freedom available for the national tests, and the smaller contrasts that result from comparing NCLB implementation levels in states where all had some level of NCLB exposure. Fortunately, the statistical results are stronger for tests that add together the weaker immediate DD and the weaker slope DD effects. Nine independent tests of differences in endpoint differences are possible across the three grade/subject matter areas. All are in the expected direction, whether the categorical or continuous measure of state proficiency is used. Of the nine, five are statistically significant at the .10 level or lower. Statistical significance is most clear in the larger sample state-level tests where all three tests attain the .10 level or lower. Two of the six tests of public versus Catholic schools also show reliable differences at this level. Thus, the overwhelming coherence in the direction of independent effects is buttressed by conventional statistical tests of the difference in differences of the *combined* mean and slope effects. So we conclude that the obtained data match the pattern that would be expected if NCLB had raised achievement.

Table 6 reports three measures of the size of the effect from 2003 to 2011. Using NAEP's suggested conversion to obtain differences in student standard deviation units, the math effects are of about .30 *SD*s averaged across all of the tests of the final endpoint differences in math, there being little difference between fourth and eighth grades. The comparable value for reading is about .10. In terms of months gained over almost a decade of NCLB, the corresponding averages are about 8.5 months in math—about a school year—and about 3 months in reading. Given the national scope of NCLB, these would seem to be large effects.

The preceding results are based on using time series data to test 39 versions of a DD interaction hypothesis—that a preintervention difference in means and slopes is different from later differences in means and/or slopes. However, the obtained results suggest a major causal contingency that elaborates the DD into DDD form. Whether Trend or Main NAEP is used, all the math coefficients are larger than the reading ones and are more likely to be statistically significant by conventional criteria. By themselves, the reading results probably have little practical policy value, though they do support the nontrivial conclusion that the NCLB activities that promoted math did not have a negative unexpected side effect on reading. Methodologically, the larger DD found with math indicates that a DDD model might have been more appropriate and would have identified the kind of boundary condition that Fisher (1935) considered a second benefit of elaborate theory; it serves to describe part of the range over which the size of a causal relationship varies.

Ruling Out Alternative Theories That Predict Some, but Not All, of the NCLB Predictions

For causal purposes, the degree of empirical coherence is only half of the story. The rest is the uniqueness of the predictions relative to contending theories. One contender is the possibility that students disproportionately left Catholic schools for public schools after 2002. This could affect means not just in Catholic schools but also private ones if (a) the students who left Catholic schools for public ones raised achievement there or (b) those who moved to non-Catholic private schools lowered means there, and (c) students leaving Catholic schools in HS states in 2002 were higher achievers than those exiting Catholic schools in LS states. All of these propositions have to be true to invalidate the hypothesis of NCLB's effectiveness. Many readers will find this concatenation of alternatives implausible in its own right, but perhaps especially in light of, first, the quite small amount of accelerated change out of Catholic schools observed in 2002—about one fifth of 1% nationally, and second, movers had no detectable impact on the post-2002 composition of public (and non-Catholic private) schools for three demographic measures correlated with achievement. Information for public schools comes from Common Core Data and for private schools from the Private School Universe Survey. Table 8 presents coefficients for the change in mean, slope, and total change by 2006 for some race/ethnic variables, some school size variables, and the student-to-teacher ratio. No reliable changes in composition were observed after 2002 and the data do not seem to the eye to be systematically changing around then. Internal validity threats based on differential compositional shifts around 2002 do not seem plausible.

Another contending theory is that school officials manipulated NAEP achievement test scores after 2002. The assumption is that they were more motivated to do this in public than private schools because NCLB only applies to the former and that they were also more motivated in states that most needed to develop higher standards. There are many ways of manipulating test scores, but two popular ones are to modify how many students with disabilities or English language learners are tested. Is there any evidence this happened around 2002, more so in HS than LS states, and more so in public than private schools? Table 9 shows the percentage of students with disabilities and English language learners taking NAEP tests immediately before and after 2002 and also for longer periods on each side of 2002. Table 10 shows the corresponding percentages of those excused. The latter reveals that more students are excluded in public than private schools after 2002, which should spuriously increase the mean change observed in these schools relative to the private ones. But no such differential exclusion pattern is evident in the HS versus LS state contrast where an NCLB effect is also claimed. Also adding to implausibility is that school officials have less motivation to manipulate the testing process for NAEP than for the state achievement tests on which consequential AYP decisions depend. The

Table 8. Difference in differences in mean, slope, and total population change post–No Child Left Behind: Contrasts of public with Catholic and with non-Catholic private school

	% Black			% Hispanic			% White		
	Coeff.	SE	t	Coeff.	SE	t	Coeff.	SE	t
Public vs. Catholic									
Diff. in mean Δ	−0.31	0.84	−0.36	0.68	0.73	0.93	−2.06	2.11	−0.98
Diff. in slope Δ	−0.25	0.25	−0.97	−0.14	0.22	−0.64	−0.50	0.64	−0.78
Diff in total Δ (2007)	−0.21	1.86	−0.11	1.91	1.60	1.20	−1.41	4.65	−0.30
Public vs. other private									
Diff. in mean Δ	4.92	12.17	0.40	−1.69	39.13	−0.04	40.05	84.19	0.48
Diff. in slope Δ	−0.39	0.35	−1.11	0.22	1.11	0.20	−1.47	2.39	−0.61
Diff in total Δ (2007)	5.70	13.74	0.42	−2.26	44.17	−0.05	45.58	95.03	0.48
	% Student			% Fourth Graders			% Eighth Graders		
Public vs. Catholic									
Diff. in mean Δ	0.55	1.30	0.42	0.25	0.45	0.56	0.11	0.78	0.13
Diff. in slope Δ	0.23	0.40	0.58	0.13	0.14	0.96	0.05	0.24	0.21
Diff in total Δ (2007)	0.27	2.42	0.11	−0.16	0.84	−0.19	−0.05	1.46	−0.03
Public vs. other private									
Diff. in mean Δ	0.09	10.93	0.01	0.71	4.95	0.14	0.23	6.48	0.04
Diff. in slope Δ	0.19	0.40	0.49	0.05	0.18	0.27	0.03	0.24	0.12
Diff in total Δ (2007)	0.00	12.32	0.00	0.72	5.58	0.13	0.19	7.30	0.03
	Student-to-Teacher Ratio								
Public vs. Catholic									
Diff. in mean Δ	0.46	0.65	0.71						
Diff. in slope Δ	0.03	0.20	0.16						
Diff in total Δ (2007)	0.46	1.44	0.32						
Public vs. other private									
Diff. in mean Δ	−5.53	34.35	−0.16						
Diff. in slope Δ	0.25	0.97	0.26						
Diff in total Δ (2007)	−6.64	38.76	−0.17						

Source: Common Core Data and Private School Universe Survey.

only manipulation-related explanation we can imagine is that high-stakes state tests have changed the culture of taking all achievement tests in ways that increase scores and that this testing change has generalized to include NAEP. If so, the NAEP-affected testing change would be more pronounced in public schools and HS states. No data exist on the plausibility of this alternative that undermines the general contention that NAEP is a test with zero stakes. It is inherently, but it has nonetheless been affected by the stakes linked to state achievement tests.

A third contending theory invokes the National Council of Teachers of Mathematics (NCTM). NCTM updated its math standards in 2000 and later claimed that this update was responsible for the subsequent national improvements in NAEP math scores (NCTM,

Table 9. Percentage of students identified with a disability and limited English proficient: By contrast group, years immediately around No Child Left Behind's (NCLB's) implementation, and averaged across all the years pre- or post-NCLB

	High Proficiency Standard States			Low Proficiency Standard States		
Year immediately before and after NCLB	*2000 or 2002*[1]	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	18.92	20.77	1.85	16.09	19.08	2.99
8th Grade Math	17.00	18.77	1.77	15.36	17.15	1.79
4th Grade Reading	19.23	21.00	1.77	18.09	18.69	0.60
Years averaged before and after NCLB	*1990–2002*	*2003–2009*	*Diff*	*1990–2002*	*2003–2009*	*Diff*
4th Grade Math	14.84	21.28	6.44	13.42	19.90	6.48
8th Grade Math	12.27	18.44	6.16	11.19	17.03	5.83
4th Grade Reading	15.37	21.51	6.14	14.17	19.79	5.62
	Public			Catholic/Non-Catholic		
Year immediately before and after NCLB	*2000 or 2002*	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	19.16	22.28	3.12	2.89	3.94	1.05
8th Grade Math	14.36	18.54	4.19	1.91	3.17	1.27
4th Grade Reading	20.56	21.91	1.35	1.52	3.20	1.68
Years averaged before and after NCLB	*1990–2002*	*2003–2009*	*Diff*	*1990–2002*	*2003–2009*	*Diff*
4th Grade Math	11.25	22.67	11.42	2.89	4.33	1.45
8th Grade Math	11.67	18.67	7.00	1.91	3.67	1.76
4th Grade Reading	16.00	22.67	6.67	2.36	4.33	1.97

[1] Closest year prior to NCLB that data is available for math (2000) and reading (2002).
Source: National Center for Educational Statistics and the Educational Testing Service.

2008). It might also explain why the current math effect was larger than the reading one in fourth grade. But for this interpretation to hold requires that NCTM standards were adopted more often, or were implemented better, in HS over LS states and in public over private schools. Maybe public schools did take the new math standards more seriously, and maybe LS states used the new standards to upgrade their standards because they had greater need of upgrading. We have found no convincing evidence on these matters across all our contrasts. But three points are worth considering. For one, a math standards explanation has to deal with the fact that a 3-year causal delay is required. For another, it is not clear that national performance can be raised merely by propagating new standards—if only national change were so easy! Finally, one evaluation concluded that every state adopted NCTM standards and did so with only minor variations, and that this did not improve math achievement and may even have harmed it (Fordham Foundation, 2005). We deem it very unlikely, but not impossible, that a delayed effect of new math standards is responsible for the complex pattern of findings reported here. Nonetheless, this conclusion requires more assumptions than a well implemented RCT or RDD would and hence greater recourse to "squishy" concepts like plausibility.

Table 10. Percentage of student with disability and limited English proficiency excluded from National Assessment of Educational Progress testing: By contrast group, years immediately around No Child Left Behind's (NCLB's) implementation, and averaged across all the years pre- or post-NCLB

	High Proficiency Standard States			Low Proficiency Standard States		
Year immediately before and after NCLB	*2000 or 2002*[1]	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	4.33	3.31	−1.03	4.09	3.85	−0.24
8th Grade Math	3.42	3.69	0.28	4.45	3.77	−0.69
4th Grade Reading	6.15	5.85	−0.31	6.91	5.85	−1.06
Years averaged before and after NCLB	*1990–2002*	*2003–2009*	*Diff*	*1990–2002*	*2003–2009*	*Diff*
4th Grade Math	6.92	3.26	−3.66	6.04	3.21	−2.84
8th Grade Math	6.06	4.03	−2.03	5.00	3.67	−1.33
4th Grade Reading	6.79	5.90	−0.89	6.17	5.92	−0.25
	Public			Catholic		
Year immediately before and after NCLB	*2000 or 2002*	*2003*	*Diff*	*2000 or 2002*	*2003*	*Diff*
4th Grade Math	4.24	3.92	−0.33	0.00	0.11	0.11
8th Grade Math	3.94	3.76	−0.18	0.00	0.04	0.04
4th Grade Reading	6.80	6.31	−0.49	0.64	0.85	0.21
Years averaged before and after NCLB	*1990–2002*	*2003–2009*	*Diff*	*1990–2002*	*2003–2009*	*Diff*
4th Grade Math	5.67	2.50	−3.17	0.00	0.15	0.15
8th Grade Math	5.00	4.00	−1.00	0.00	0.18	0.18
4th Grade Reading	6.50	6.33	−0.17	0.64	0.64	0.00

[1] Closest year prior to NCLB that data is available for math (2000) and reading (2002).

Source: National Center for Educational Statistics and the Educational Testing Service.

CONCLUSIONS

How Adding Multiple Design Elements to a One-Group ITS Design Improves Internal Validity

This article seeks to illustrate the utility of augmenting a short interrupted ITS when an RCT or RDD is not possible. Shadish et al. (2002) outlined the internal validity threats most often linked to simple ITS studies, and we use them to assess how adding other design elements helped achieve a stronger causal inference with NCLB.

Statistical regression is not an issue, because there is no visual evidence of any sudden shifts in achievement means just before NCLB was introduced. Some of the pre-NCLB data series ended several years before 2002, but others ended in 2002 and would be especially likely to manifest a dip if there was one. But there was none. A change in *instrumentation* immediately after 2002 is not an issue either. No item changes occurred with Trend NAEP, and it showed initial statistically significant changes in both fourth- and eighth-grade math means. Moreover, it will be difficult to come up with accounts of why Main NAEP test items or testing procedures were suddenly implemented differently after 2002 and why this new implementation differed between both public and private schools nationally as well as between HS and LS states. A similar argument applies to *maturation* threats. Although the supplemented DD analysis of data from a basic ITS design takes care of different linear maturation rates in the treatment and comparison group, the more remote possibility exists of a inflection in the maturation rate occurring differentially in 2002 and thus invalidating an NCLB causal effect. On top of this, after 2002 maturation rates would also have to suddenly change in a differential way for HS versus LS states and for public versus private schools of two kinds.

A specific differential *selection* change after 2002 is also not plausible. Theoretically, news about sexually abusive priests around 2002 might account for the obtained data in the national contrast of public and Catholic schools. However, Table 1 showed that that any post-2002 acceleration in the rate of enrollment loss in Catholic schools is about 0.20% of all national students. More important, there is no evidence that enrollment changed after 2002 in non-Catholic private schools. Yet effect sizes were the same whether Catholic or non-Catholic private schools provided the contrast; the accelerated loss of students in Catholic schools after 2002 could affect the contrast only between HS and LS states if it were also the case that more, or different kinds of, students left Catholic schools in the LS states than in the HS ones. This is possible but, we judge, hardly probable.

What about *history* as an alternative interpretation? Adding one or more comparison groups results in a DD approach that rules out all national historical events around 2002 that equally affected both HS versus LS states and public versus private schools. That still leaves the possibility of historical events after 2002 that differentially affected each of these contrasts. It is not easy to conceive of such events, but the recommended change in math standards in 2000 is one possibility, provided one is willing to assume a 3-year causal delay period for which there is no warrant in previous research and one can invoke a plausible rationale for why issuing new standards is a strong enough procedure to bring about math gains in the nation at large! Of course, math standards might have affected achievement because NCLB led public schools and HS states to cast around more actively for new ideas about teaching math. But this account treats the 2000 math standards as a consequence of NCLB rather than as a threat to internal validity.

What about data manipulation by school officials as an alternative interpretation? Although this has been repeatedly observed with state achievement tests, the outcome here is NAEP. Seemingly, it entails no consequences for schools and so should arouse little incentive to manipulate scores. However, it is possible that the entire culture of achievement tests has changed in schools and that score-inflating testing or review procedures are now built into all achievement testing. If so, the pressures might be stronger in public than private schools and in some states more than others, in line with the different consequences that state test-taking entails for them. This possibility is hypothetical, and we know of no data concerning whether schools experienced NAEP testing differently after 2002 and in a way systematically mirroring the pattern of data obtained here.

Each of the design elements added to the basic ITS structure has a role in ruling out alternative interpretations. The Trend NAEP data help rule out an instrumentation alternative interpretation. The replication of results across both fourth- and eighth-grade math help rule out chance and increases external validity by showing that the math effect is not specific to a single grade. The use of multiple contrasts helps generate a national estimate for a national program like NCLB, and the state contrast helps rule out a selection alternative interpretation predicated on an accelerated loss of enrollment in Catholic schools after 2002—as does also adding the non-Catholic private comparison group. Indeed, the three comparison groups rule out most history alternatives, as it is unlikely that the same difference in math-related local historical events took place between states with different proficiency standards and between public and private schools. No single comparison time-series could rule out all these different interpretations; multiple comparison series were needed.

However, adding multiple design elements is not a panacea. Little is gained by multiple hypothesis tests if they leave out a plausible internal validity threat or if they replicate the same source (or direction) of bias. Collectively, the different comparison series, datasets, and outcomes have to produce a pattern of obtained effect estimates that is coherent with the initial pattern of expected effects *and that is capable of ruling out all plausible threats to internal validity* (Cook, 1985). Pattern matching allows for a more severe test than a single ITS, understanding severity as a high probability of detecting an error when present and a low probability of indicating an error when absent (Mayo, 1996; Popper, 1983). But coherence or pattern matching does not guarantee a perfect causal test, and we were not able to definitively rule out convoluted versions of a manipulation or belated math standards effect without recourse to arguments outside of the time series or other data considered here.

Would a Simpler Design Have Been More Effective?

We have claimed that the multiply supplemented ITS design used to evaluate NCLB has led to plausible estimates of the program's national impact. But would a simpler design have been just as effective? In this regard, consider three possible alternatives. First is an evaluation of NCLB that used administrative waivers to facilitate random assignment. The political problem here is getting such waivers; they run contrary to the spirit of a new national law and were not used with NCLB. Moreover, if such an RCT could happen its causal estimates would only generalize to entities that applied for a waiver and met any restrictions imposed on those who could receive a waiver. In contrast, the ITS study includes the nationally representative sample of public schools used for NAEP testing and so has greater external validity.

Second, consider an RDD study that takes advantage of state cutoffs for making AYP. The difficulties here are that NCLB has 39 different cutoff values for different subgroups and subject matters, and there is considerable evidence of deliberate manipulation of scores around the cutoff in ways that preclude recreating the score actually obtained on the assignment variable (Wong, 2011). The 39 cutoff values complicate RDD analysis considerably; and the possibility of deliberate manipulation of scores entails possible bias. Moreover, RDD involves external validity limitations too, as causal estimates are only warranted at the cutoff unless the design includes a pretest or comparison RD function.

Finally, consider implementing a matching design instead of an RCT or RDD. Matching is a large sample technique and so is not possible with states as the unit; schools are more appropriate. Using restricted use data tapes, public schools could indeed be matched to Catholic and non-Catholic private schools without recourse to an ITS framework, as can schools in HS and LS states. But public and private schools are characterized by different achievement means, making it likely that some schools in each population could not be matched and would have to be omitted, thus reducing external validity relative to our ITS design that compares representative samples of *all* public and *all* private schools. Also, achievement means are higher in private than public schools, and so the lowest scoring public schools are least likely to be matched—the very public schools NCLB was designed to help most. We could also match schools within states that vary in how much change NCLB requires of them. But again we might lose some schools because of the population difference between states. Matching becomes even more problematic if we aspire to preserve the time-series structure of the data for a DD approach. In each NAEP assessment cycle, a different sample of schools and students is selected, thus precluding a genuinely longitudinal school-level matching. Instead, treatment and control schools would need to be matched separately by year, entailing different school populations by year that will introduce error and perhaps bias.

The national ITS estimates we have presented generalize to the nation at large. Any marginal internal validity gains that an RCT, RDD, or sophisticated matching design would confer has to be judged against the reduced external validity it might entail. We have no way to quantify the trade-offs in question, but it is not intuitively clear to us that any of the alternative designs promising higher internal validity would continue to be superior if internal and external validity were *jointly* taken into account in assessing design quality. The weak case for a multiply supplemented ITS design is that it is a viable option for evaluating national and statewide programs when better causal designs are not possible. The stronger (but not easily quantified) case is that multiply supplemented designs may not be inferior to RCTs and RDDs *if internal and external validity are jointly used to assess design quality*.

The methodological exercise we have just presented adds to prior knowledge of NCLB's effects.

1. We now have national estimates of the effects of NCLB by 2011.
2. We now know that NCLB affected eighth-grade math, something not statistically confirmed in either Wong, Cook, Barnett, and Jung (2008) or Dee and Jacob (2011) where positive findings were limited to fourth-grade math.
3. We now have consistent but statistically weak evidence of a possible, but distinctly smaller, fourth-grade reading effect.
4. Although it is not clear why NCLB affected achievement, some possibilities are now indicated.

One is what Dee and Jacob (2011) call "consequential accountability"—sanctions for which the likelihood of being implemented varies by state. Another is what we have called "higher state standards" that engender different amounts of educational reform because of how NCLB's regulations are framed. A third possibility combines both mechanisms, as there is no correlation between the states with high standards and those that strictly apply their sanctions ($r = .05$). As NCLB's 2014 end-date for universal proficiency approaches, more and more schools are failing to make AYP and district, state and federal officials face mounting political and financial pressure to modify the program in fundamental ways. By 2013, more than half of all states have been granted waivers, and Congress is actively considering discontinuing the use of state tests to determine school failure. Yet NCLB has reduced math achievement gaps between public and private schools, and these two kinds of schools differ in their students' income and race backgrounds.

FUNDING

This work was supported by the Institute of Education Sciences' Grants R305U07003 and R305D100033. Special thanks go to Christopher Jencks, Larry V. Hedges, and Vivian Wong for feedback.

REFERENCES

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Belasen, A. R., & Dai, C. (2011). *When oceans attack: Using a generalized difference-in-difference technique to assess the impact of hurricanes on localized taxable sales*. Unpublished manuscript, Southern Illinois University, Carbondale, IL.

Belasen, A. R., & Polachek, S. W. (2008). How hurricanes affect wages and employment in local Labor Markets. *The American Economic Review*, *98*(2), 49–53.

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik*. New York, NY: Holt, Rinehart, & Winston.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. *Compensatory Education: A National Debate*, *3*, 185–210.

Center on Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington, DC: Author.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, *99*, 1145–1177.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society. Series A 128*, 134–155.

Cook, T. D. (1985). Post-positivist critical multiplism. In R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21–62). Beverly Hills, CA: Sage.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology*, *45*, 545–579.

Corrin, W. J., & Cook, T. D. (1998). Design elements of quasi-experiments. In A. J. Reynolds & H. J. Walberg (Eds.), *Advances in educational productivity* (Vol. 7; pp. 91–112). Greenwich, CT: JAI Press.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*, 418–446.

Ewing, B. T., & Kruse, J. B. (2005). *Hurricanes and unemployment (*Center *for Natural Hazards Research)*. Greenville, NC: East Carolina University.

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low performing schools. *American Educational Research Journal*, *44*, 594–630.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Fordham Foundation. (2005). *The state of State math standards*. Washington, DC: Author.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind. *Educational Researcher*, *36*, 268–278.

Goertz, M. E. (2005). Implementing the No Child Left Behind Act: Challenges for the states. *Peabody Journal of Education*, *80*, 73–89.

Hill, C., Bloom, H., Black, A., & Lipsey, M. (2007). *Empirical benchmarks for interpreting effect sizes in research.* New York, NY: Manpower Demonstration Research Corporation.

Imbens, G., & Wooldridge, J. (2007). *What's new in econometrics (*Lecture *notes)*. NBER Summer Institute, Cambridge, MA.

Keigher, A. (2009). *Characteristics of public, private, and Bureau of Indian Education elementary and secondary Schools in the United States: Results from the 2007–08 Schools and Staffing Survey (*NCES *2009–321)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Lazer, S. (2004). Innovations in instrumentation and dissemination. In L. Jones & I. Olkin (Eds.), *The Nation's report card: Evolution and perspectives* (pp. 469–487). Bloomington, IN: Phi Delta Kappa Educational Foundation and American Educational Research Association.

Lechner, M. (2010). The estimation of causal effects by difference-in-difference methods. *Foundation and Trends in Econometrics*, *4*, 165–224.

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into National and State reading and math outcome trends*. Cambridge, MA: Harvard Civil Rights Project.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge.* Chicago, IL: The University of Chicago Press.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, *13*, 151–161.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *The Review of Economics and Statistics*, *86*, 156–179.

Milligan, K., & Stabile, M. (2009). Child benefits, maternal employment, and children's health: Evidence from Canadian child benefit expansions. *American Economic Review*, *99*, 128–132.

National Center for Education Statistics. (2009). *How the samples of schools and students are selected for the main assessments (*State *and National).* Retrieved from http://nces.ed.gov/nationsreportcard/about/nathow.asp

National Council of Teachers of Mathematics. (2008, January/February). Rise in NAEP math scores coincides with NCTM standards. *NCTM News Bulletin*, pp. 1–2.

Neal, D., & Schanzenbach, D.W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*, 263–283.

Popper, K. R. (1983). *Postscript: Vol. 1. Realism and the aim of science*. Totowa, NJ: Rowman & Littlefield.

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin*, *13*, 19–23.

Rosenbaum, P. R. (2005). Observational studies. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1451–1462). Chichester, UK: Wiley & Sons.

Rosenbaum, P. R. (2009). *Observational studies (*Edition *2).* New York, NY: Springer-Verlag.

Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, *106*, 285–295.

Rosenberg, M. S., Sindelar, P. T., & Hardman, M. L. (2004). Preparing highly qualified teachers for students with emotional or behavioral disorders: The impact of NCLB and IDEA. *Behavioral Disorders*, *29*, 266–278.

Shadish, W., Cook, T. D., & Campbell, D. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton Mifflin.

Smith, T. M., Desimone, L. M., & Ueno, K. (2005). Highly qualified to do what? The relationship between NCLB teacher quality mandates and the use of reform-oriented instruction in middle school mathematics. *Educational Evaluation and Policy Analysis*, *20*, 75–109.

Solé-Ollé, A., & Sorribas-Navarro, P. (2008). The effects of partisan alignment on the allocation of intergovernmental transfers. Differences-in-differences estimates for Spain. *Journal of Public Economics*, *92*, 2302–2319.

Stullich, S., Eisner, E., & McCrary, J. (2007). *National assessment of Title I: Final report*. Washington, DC: Department of Education.

U. S. Department of Education. (2007). *Private school participants in Federal program under the No Child Left Behind Act and the Individuals with Disabilities Education Act*. Washington, DC: Author.

U. S. Department of Education. (2008). No Child Left Behind Act of 2001, Public Law print of PL 107–110. Retrieved from http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

Wong, V. C. (2011). *Games that schools play: Manipulation of the assignment mechanism by schools under No Child Left Behind.* Washington, DC: Society for Research on Educational Effectiveness.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness based evaluation of five state pre kindergarten programs. *Journal of Policy Analysis and Management*, *27*, 122–154.